Available online at www.sciencedirect.com

SciVerse ScienceDirect

Physics Procedia 33 (2012) 1248 – 1253

Physics Procedia

2012 International Conference on Medical Physics and Biomedical Engineering

An Algorithm of Mining Association Rules Based on Granular Computing

Xiaojun Cao

The Information Engineering School of Lanzhou University of Finance and Economics
Lanzhou, China
Caoxj@lzcc.edu.cn

Abstract

In view of the defects of Apriori association rule mining algorithm needed to scan database frequently and generate a set of large candidate. In this paper, the binary string was used to express information granule, using grain-bit binary extraction frequent item sets and discovering association rules. Through experiments, the classical Apriori algorithm and the algorithm that based on granular computing association rules extraction were compared to the experimental results are analyzed. The results show that the algorithm based on granular computing for mining association rules is feasible and effective.

*Keywords:*Association rule; Apriori algorithm; Granular computing ;Grain-bit binary extraction

1. Introduction

As a classical association rule, the Apriori association rule mining algorithm has many problems as follows[1-2]:

1)Needing to scan the database repeatedly ;

2) Generating lots of candidate item sets .

To solve the problem above, the paper's mining is from the perspective of association rule mining. In this paper, improving the algorithm in four areas as follows:

1) Reducing the I / O time by reducing the times of scanning the database.

2) Calculating the candidate itemsets as soon as possible.

3) Using a subset of candidate itemsets to get the greatest degree of decomposition.

4) Generating candidate itemsets parallelly .

 doi:10.1016/j.phpro.2012.05.206

From the perspective of particle size for mining association rules, the superset of frequent patterns can be seen as a "coarse", so the sub-frequent patterns can be seen as "fine", then the frequent patterns of parent-child relationship can be seen as a particle size space a kind of partial order.

This paper from another angle of granular computing ideas for association rule mining using, using two-dimensional table to store information, massing the data by appropriate granulation and the synthesis and decomposition of particles to reduce the complexity of the problem solving and improving the computational efficiency, at last, the simulation results prove the feasibility.

2. GRANULAR COMPUTING

Granular computingbasic problem consists of two aspects, firstly, how to build information granularity, the other is how to use granular to compute. The purpose of granular computing is try to find the smallest computational complexity approximate solutions satisfied enough feasibility within the scope of allowable error. [3]

2.1 Granular Computing

Computing depends on the previously discussed notion of granulations. They can be similarly studied from both the semantic and algorithmic perspectives. The two level structures, the granule level and the granulation level, provide the inherent relationships that can be explored in problem solving. The granulated view summarizes available information and knowledge about the universe. As a basic task of granular computing, one can examine and explore further relationships between granules at a lower level, and relationships between granulations at a higher level.

The relationships include closeness, dependency, and association of granules and granulations. Such relationships may not hold fully and certain measures can be employed to quantify the degree to which the relationships hold. This allows the possibility to extract, analyze and organize information and knowledge through relationships between granules and between granulations [4-5]

The problem of computing and reasoning with granules is domain and application dependent. Some general domain independent principles and issues are listed below.

• Mappings between different level of granulations. In the granulation hierarchy, the connections between different levels of granulations can be described by mappings.

• Granularity conversion. A basic task of granular computing is to change views with respect to different levels of granularity.

• Property preservation. Granulation allows the different representations of the same problem in different levels of details..

• Operators. The relationship between granules at different levels and conversion of granularity can be precisely defined by operators.

Granular computing methods describe our ability to switch among different granularities in problem solving. For example, concrete domain specific conversion methods and operators can be defined. In spite of the differences between various methods, they are all governed by the same underlying principles of granular computing.

2.2 Association Rules

Using association rules for dealing the data mining.The following is a formal statement of the problem:Let L={i1,i2,...,im} be a set of literals, called items.Let D be a set of transactions,where each transaction T is a set of items such that $T \subseteq L$. Associated with each tranction is a unique identifier, called

TID. We say that a transaction T contains X, a set of some items in L, if X ⊆T. An asstion rule is an implication of the form X ==> Y ,where X⊆ L,Y ⊆L,and X ==>Y= ∅ .The rule X ==>Y holds in the transaction set D with confidence c if c% of transaction in D that contain X also contain Y. The rule X ==>Y has support s in e the transaction sed D if s% of transaction in D contain X∪Y .

3、RELATED WORK

3.1 The Mining Algorithm Based on Granular Computing Association Rule(AGC_AR)

Algorithm idea: The mining algorithm based on Granular computing association rule(AGC_AR) first used the granular computing theory to decompose the information system, then logical operations elementary particles in accordance with the requirements, obtaining the frequent itemsets. The collection process of Particle element method as follows:

Input: An information system S = (U, C ∪ D, V, f).

Output: Information systems's space and the set of elements Qgs (Gri) of all particles Qgrs

1) Order the Qgrs = ∅, m = | Attr |, n = | U |.

2) Diving each attribute in turn divided by an equivalence relation: U / IND (Attri), ∀ qgij ∈ U / IND (Attri) is the first i attributes of the first j equivalence class.

3) Scanning the database, obtaining the support of each qgij the set of elements and particles qgs (g).

4) For any particle element qgij, if its support is greater than the minimum support minSup, adding it into the particle collection QgrS.

Algorithm1 is computing information granules's support , recording a collection of particle elements at the same time, which made different items at the beginning of the item set can be solved for different set of objects.

Then using the algorithm 2 for getting all the frequent itemsets to reduce the scanned object.

Input: Particle space Grs.

Output: All the frequent itemsets Gfrs

1) Order Gfrs = ∅.

2)There are Gfrs. | Gfrs | items tablets in the total commercial space, beginning itemsets ItemSets (Gri) capsules of each element one by one to survive as Gri, while elements of its particle collection qgs (Gri) items within the set of demand support, if a subset has been identified as not frequent, which is no need to calculate their support.

3)For the itemSets (Gri) in any one, if its support is greater than minsup,then it is frequent, joining the frequent items in the collection GfrS. All its subsets are frequent, therefore, all joined GfrS in the subset. of particles carried out in order to reduce the problem space

The main reason of algorithm scaned the database is for obtaining the support of each candidate

3.2 An example of algorithm'use:

There is a transaction database, which has nine service numbers,two of them are the transaction number and the contents of storage services.The support is 2, the table as below::

TABLE I .TRANSACTION DATABASE

TID	GOODS
001	BDE
002	AD
003	CD
004	ABD
005	BC
006	CD
007	BC
008	BCDE
009	BCD

First using the Apriori algorithm for mining of association rules to the above data, the specific process as follows: Obtaining a set of processes frequently as shown in the table:

TABLE Ⅱ SET OF ONE FREQUENT

Itemset	Support
A	2
B	6
C	6
D	7
E	2

In the above table, scanning the database for each item of the set and counting it, removing those whose support is less than the key support, as all those in a concentration is greater than the support, so all retain.,as the following :

TABⅢ SET OF TWO FREQUENT

Itemset	Support
{A,D}	2
{B,C}	4
{B,D}	4
{B,E}	2
{C,D}	4
{D,E}	2

In the above table, comparing candidate support count and the minimum support, removing those does not meet the minimum support required.,for example, (B, C) can connect with (B, D) but do not with (C, D), with this principle, getting (B, C, D), (B, C, E), (B, D, E), because (C, E) is not a frequent subset of (B, C, E), so deleting (B, C, E) on these three items , which also can not be frequent.. As is shown in the table.：

TABⅣ SET OF THREE FREQUENT

Itemset	Support
{B,C,D}	2
{B,D,E}	2

After using the same method to connect and pruning, because it does not meet the join condition so the next candidate set is empty,and the Apriori algorithm is over.

Second using the algorithm based on granular computing association to the above data The mining association rules under the two-dimensional tables to store information tablets, firstly scanning the database to create basic ,as is shown in the table.

TAB Ⅴ THE TABLE OF GRANULAR COMPUTING'S BINARY

Grain	Information Granules	Binary System	Grains size
[A]	{002,004}	010100000	2
[B]	{001,004,005,007,008,009}	100110111	6
[C]	{003,005,006,007,008,009}	001011111	6
[D]	{001,002,003,004,006,008,009}	111101011	7
[E]	{001,008}	100000010	2

In the above table, all of the size to meet the minimum support, so all of them are often a set of content. After we get the frequent set period, setting all of the grain to grain in combination to generate candidate 2 sets, so we step through the synthesis of a domain can be the combination tablets name: [A, B], [A, C], [A, D], [A, E], [B, C], [B, D], [B, E], [C, D], [C, E], [D, E] . The binary operation is a binary combination with tablets of intersection, the result as shown in the table.

TAB Ⅵ SET OF TWO FREQUENT OF GRANUALR COMPUTING

Grain	Information Granules	Binary System	Grains size
[A,D]	{002,004}	010100000	2
[B,C]	{005,007,008,009}	000010111	4
[B,D]	{001,004,008,009}	100100011	4
[B,E]	{001,008}	100000010	2
[C,D]	{003,006,008,009}	001001011	4
[D,E]	{001,008}	100000010	2

Obtaining the domain of candidate itemsets synthesis by the new combination tablets,focusing on the frequent 2 to connect three new particles:[B, C, D], [B, C, E], [B, D , E].By the false principles of commercial space, [C, E] is non-candidate frequent set, so only [B, C, D], [B, D, E] are frequent sets,but [B, C, E] is not. While the binary representation of their binary and computing the results as below:：

TAB Ⅶ SET OF THREE FREQUENT OF GRANULAR COMPUTING

Grain	Information Granules	Binary System	Grains size
[B,C,D]	{008,009}	000000011	2
[B,D,E]	{001,008}	100000010	2

From the above algorithm,finding that for the classical Apriori algorithm, which needs to scan the database many times to statistics the number of candidate itemsets and need many pattern matching when scanning the database.That will spend a lot of time and produce a large number of of candidate itemsets,which also take the bottleneck, but the mining algorithm based on Granular computing association rule has no this question. The comparison of two algorithms's execution time as shown:

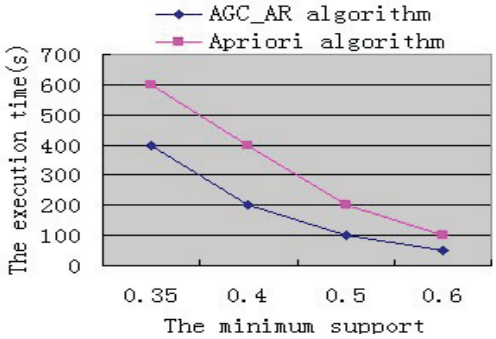

Fig.1. The comparison of algorithm's execution time

4.CONCLUSION

With the rapid development of the network and database technology, requiring increasing scale data to be processed, therefore, how to data mining effectively for pair of massive data is a serious problem.Maturing granular computing algorithm has provided new methods and ideas for data mining research. This particle size calculated of the association rule mining consider the frequent item sets, a collection of objects to reduce the scanning itemsets, improving the efficiency of the algorithm, and the test proved that this algorithm is valid for data volume.

REFERENCES

[1] Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., and Verkamo, A.I. 1996. Fast discovery of association rules.In Advances in Knowledge Discovery and Data Mining, U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R.Uthurusamy (Eds.), AAAI/MIT Press, pp. 307–328.

[2] C. Borgelt and R. Kruse. Induction of Association Rules:Apriori Implementation. Proc. 14th Conf. on Computational Statistics (COMPSTAT). Berlin, Germany 2002.

[3] Y.Y. Yao. Information granulation and rough set approximation,International Journal of Intelligent Systems, Vol. 16, No. 1, 87-104, 2001.

[4] Y.Y. Yao.A Partition Model of Granular Computing.Department of Computer Science, University of Regina, Saskatchewan, Canada. http://www.cs.uregina.ca/↑ yyao.

[5] S. Kramer, L. de Raedt, and C. Helma. Molecular Feature.Mining in HIV Data. Proc. 7th Int. Conf. on Knowledge Discovery and Data Mining (KDD-2001, San Francisco, CA),136–143. ACM Press, New York, NY, USA 2001